International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN (P): 2249-6831; ISSN (E): 2249-7943
Vol. 7, Issue 5, Oct 2017, 29-34

INSIGHT OF VARIOUS POS TAGGING TECHNIQUES FOR HINDI LANGUAGE

SIMPAL JAIN[1] & NIDHI MISHRA[2]

[1]*M.Tech Scholar, Poornima University, Jaipur, India*

[2]*Associate Professor, Poornima University, Jaipur, India*

ABSTRACT

Natural language processing (NLP), is the process of extracting meaningful information from natural language. Part of speech (POS) tagging is considered as one of the important tools, for Natural language processing. Part of speech is a process of assigning a tag to every word in the sentences, as a particular part of speech, such as Noun, pronoun, adjective, verb, adverb, preposition, conjunction etc. Hindi is a natural language, so there is a need to perform natural language processing on Hindi sentence. This paper discussed a hybrid based approach, for POS tagging on Hindi corpus. This paper discussed a review of different Techniques, for Part of Speech tagging of Hindi language.

KEYWORDS: Hidden Markov Model, POS Tagging, Hindi Word Net & Hybrid.

Received: Aug 20, 2017; **Accepted**: Sep 17, 2017; **Published**: Oct 13, 2017; **Paper Id**.: IJCSEITROCT20173

Original Article

INTRODUCTION

Natural language processing is a broad area of computer science and artificial intelligence. Part of speech is a very important application for NLP. A sentence is made of words, which play their different part in the framework of the sentence. Words can broadly be classified, on the basis of the part they play, or work they do in a sentence.

These are called the Parts of Speech (POS) which are noun, conjunction, adjective, adverb, preposition, pronoun, verb, etc. Ambiguity across POS categories is the biggest challenge in Part of Speech, where a word has got multiple tags in the post categories. For example "सोने"" can be treated as a noun or verb. Hindi POST is the process of identifying the lexical category of the Hindi word, existing in a sentence. [3] Part of Speech tagging can be done, using many techniques, i.e. Rule based, stochastic (or Statistical) and Hybrid.

Natural Languages are ambiguous in nature. At different levels of Natural language processing (NLP), task ambiguity appears. Multiple part of speech tags are taken by many words. The correct Tag depends on the context. [4]

For Example

भारत	सोने	की	चिड़िया	हैं
NN	VM/	PSP	NN	VM
	NN			

Figure 1: POS Ambiguity of a Hindi Sentence with Seven Basic Tags

In figure 1 the word "सोने" can be a verb or can be a Noun. [4]

LITERATURE SURVEY

Many researches are carried out in POS tagging for Hindi languages. There have many implementations using Rule Based approach, Statistical approach and Hybrid Approach. Hybrid approach provides higher accuracy, compared to rule based and statistical.

Nidhi Mishra, et-al, 2011, proposed Part of Speech Tagging for Hindi Corpus. The system implemented a Hindi corpus of 4 lines, 7 sentences and 68 words. They split the sentences into words, using space delimiter, and then assigned a particular part of speech to each Hindi word such as Noun, Pronoun, Verb, Adjective etc. They also displayed a tag structure and corresponding sentence in the grid, according to tag pattern. [1]

Sanjeev Kumar Sharma, et-al, 2011, proposed a Panjabi POS tagger, using A Bi-gram Hidden Markov Model. Author used Viterby algorithm, to implement the HMM approach. This module has been tested on a corpus of 26,479 words. The achieved accuracy of the system is 90.11% [10]

Shubhangi Rathod, et-al, 2015, discussed different POS tagging Techniques, for Indian regional language. They discussed Rule based, statistical and hybrid approach. [2]

Dilmi Gunasekara1, et-al, 2016, developed a POS tagger, using hybrid approach for Sinhala Language. Firstly, they used the HMM approach as a statistical approach. Author used stemmer to increase the accuracy. Then, author used rule based approach to assign relevant tag to the word. The achieved accuracy of the system is 72%. **[11]**

Kanak Mohnot, et-al, 2014, proposed Hindi Part of speech tagger, using Hybrid approach. Firstly, author enters a Hindi corpus and then tokenize Hindi corpus into sentences, using delimiter like "? |, !". Then, select a sentence and tokenize it into words, using space delimiter. It uses a Hindi World Net dictionary and assigns a tag to every word, occurring in the sentences. If there is a word, which is not tagged using Hindi WordNet, then it applies rule based approach to tag all words. It removes the ambiguity, using the HMM approach as a statistical approach. The accuracy achieved by the system was 89.9%. [3]

Navneet Garg, et-al, 2012, proposed Rule Based Part of Speech Tagger for Hindi. At the first phase, tag is found in the database. If it is not found in the database, then author applied various rules to tag the sentences. The system is evaluated using a corpus of 26,149 words. The achieved accuracy was 87.55 %. [4]

Pravesh Kumar Dwivedi, et-al, 2015, developed a Hindi POS tagger, using Hybrid approach. The system is evaluated using a corpus of 500 sentences. [7]

Abhijit Paul, et-al, 2015, proposed POS tagging for Nepali language, using HMM approach as a statistical approach. In this author used Nepali corpus, which contains 1, 50,839 words. The achieved accuracy was 96% of known words, but achieved less accuracy for unknown words. [6]

Antony P J, et-al, 2011, discussed various POS tagging Approaches, to assign tags for Indian Language. This paper presented a review of the various developments of POS tagger. [8]

Shachi Mall, et-al, 2015, proposed four different algorithms for Hindi POS tagging. Author Implement a corpus of 300 Hindi sentences. Firstly, author used tokenize algorithm to tokenize the Hindi paragraph and apply some rules. Achievable accuracy was 92.4%. Then author used a conversion algorithm, which translated the Hindi word into English transliteration word. Achieved accuracy was 95.7%. Third algorithm is for POS tagging, Achieved accuracy was 95.5%.

Forth algorithm is a translation algorithm, to convert the grammatical tag word into English Tagging. Accurately, the label is 95.5%. Forth algorithm is a translation algorithm, to convert the grammatical tag word into English translation, by using with Hindi to English dictionary. Accurately, the label is 96.7%. [9]

Table 1

Proposed System	Technology Used	No of Words	Accuracy	Remark
Panjabi POS tagger	Hidden Markov Model, Viterby algorithm	20,000 words	90.11	Proposed system didn't perform well due to the data sparseness problem of Panjabi.
POS tagging for Sinhala language	Hybrid	100,917	72.14%	Hybrid approach gave a higher accuracy for Sinhala language.
Hindi POS tagger	Hybrid	NA	89.9%.	The proposed system achieved high accuracy.
Hindi POS tagger	Hybrid	26,149 words	87.55 %.	Rule based POS tagger provide less accuracy compare to Hybrid approach.
Nepali POS tagger	Statistical approach	1,50,839 words	97 % of known words 43% of unknown words	The proposed POS doesn't perform well for Unknown words.

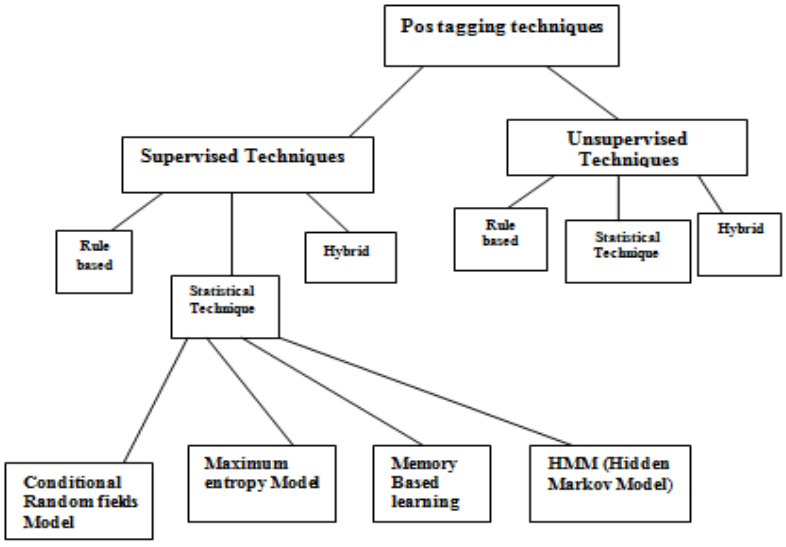

Figure 2: Classification of POS Tag Techniques

POS TAGGING TECHNIQUES

POS tagging techniques can be categorized into two approaches:

- Supervised.
- Unsupervised

Supervised

Supervised POS tagger uses pre tagged corpora. It is used to develop any tool, which will be used for tagging process. For ex: The tagger dictionary, a set of rules etc.

Unsupervised

Unsupervised POS tagger does not use pre tagged corpora, while they use advanced computational techniques to automatically make tag sets. For ex:

Baum-Welch algorithm is used to make tag sets. Again supervised and unsupervised techniques are fallen into three subcategories.

- Rule based
- Stochastic or Statistical based POS tagger
- Hybrid

Rule Based POS Tagger

Rule based POs tagger apply a set of Hand written rules, to resolve the tag Ambiguity. Rules are written on the basis of next and previous tags. It also uses contextual information, to assign tags to words in rule based tagging. It needs expressive rules and requires good knowledge of grammar related rules. [3]

For example

Rule 1

If a present word is Postposition (PSP), then there will be a high probability that the next word is a noun (NN).

For ex: राम ने खाना खाया |

Rule 2

If a present word is an adjective (Adj)

Then, there will be a high probability that the next word is a noun (NN).

For ex: सीता को कच्चा आम पसन्द हैं |

Stochastic or Statistical Based POs Tagger

The stochastic POS tagger is based on the probabilities of occurrences of words for a particular tag.

Stochastic base POS tagger can be implemented using four Models:

- Conditional Random Fields
- Maximum entropy Model
- Memory based learning
- Hidden Markov model

Conditional Random Fields

CRF (Conditional random fields), is a statistical modeling method. It is a probabilistic method, used for structure prediction. CRF is a type of discriminating undirected probabilistic graphical model, which defines a single exponential model. The benefit of CRF over hidden Markov model (HMM) is conditional nature, i.e., it doesn't require independence assumption. The advantage over MEMM (Maximum Entropy Markov Model), is the avoidance of label bias problem of MEMM. [3]

Maximum Entropy Markov (MEM) Model

MEM (Maximum Entropy Markov) model or conditional Markov model, is a graphical sequence model, that combines features of hidden Markov models (HMMs) and maximum entropy (Max Ent) models. It can represent different features of a word and can also deal with long term dependency. It uses the principle of maximum entropy. This principle states that, the least biased model is the one which maximize entropy. This model considers all the known facts, to maximize entropy. The advantage of MEMM over HMM is dealing with diverse and overlapping features. The label bias problem is the disadvantage of this approach. [3].

Hidden Markov Model

HMM is a stochastic (statistical) approach. It is a probabilistic model. HMM based POS tagger, calculates the forward and backward probability of tags, along with the input sequence, and assigns the best tag to a word. [4]

The following equation is used to assign best tag:

P(ti/wi)=P(ti/ti-1).P (ti+1/ti).P(wi/ti)

P (ti/ti-1) is the probability of present tag given previous tag.

P (ti+1/ti) is the probability of future tag given present tag.

P (wi/ti) is the Probability of word given present tag.

To compute these probabilities the following equation is used:

$$\mathbf{P\ (ti/ti\text{-}1)} = \frac{freq(ti-1,ti)}{freq(ti-1)}$$

To calculate Each tag transition probability count, the occurrences of two tags which are seen together in the corpus and divide it by the no. of occurrences of the previous tag, which are seen independently in the corpus. [4]

TABLE I. COMPARISON OF POS TAGGING APPROACHES

POS Tagging Approaches	Rule Based	Statistical	Hybrid
Description	It applies a set of hand written rules.	It is based on the probabilities of occurrences of words for a particular tag.	It is a combination of rule based and Statistical approach
Strengths	It uses a small and simple rule set.	More accurate compared to rule based tagger.	Higher accuracy compared to an individual rule based POS tagger or stochastic POS tagger.
Weaknesses	Less accurate		For an unknown word, it does

	compared to Statistical POS tagger		not assign a correct tag.

Hybrid POS Tagger

It is a combination of Rule based and stochastic based POS tagger. In this, the most probable tag is assigned to the word, using the stochastic based POS tagger. If a tag is wrong, then ruled based POS tagger is applied. [3]

CONCLUSIONS

The Hindi Word Net is a rich resource, it is being used by many Hindi Natural language processing (NLP) applications. Hindi WordNet consists of around 1 lakh unique class category of words like Noun, verb, adjective, and adverb. But still, many words are not tagged, so we use Rule based approach to assign tags to all words, and use context rules to disambiguate stochastic based approach, assigns the most likely tag to a word, based on the on-set values frequency in a corpus. Hybrid based tagging, is a combination of the two approaches. We concluded that, Hybrid Approach provides higher accuracy, as compared to an individual rule based POS tagger and stochastic POS tagger.

REFERENCES

1. *N. Mishra and A. Mishra, "Part of Speech Tagging for Hindi Corpus,"* 2011 International Conference on Communication Systems and Network Technologies, *Katra, Jammu, 2011*
2. *Shubhangi Rathod and Sharvari Govilkar," Survey of various POS tagging techniques for Indian regional languages",2015 International Journal of Computer Science and Information Technologies,2015*
3. *Kanak Mohnot, Neha Bansal, Shashi Pal Singh, Ajai Kumar" Hybrid approach for Part of Speech Tagger for Hindi language", 2014 International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2014*
4. *Garg, N., Goyal, V., Preet, S.: Rule based Hindi part of speech tagger. In: Proceedings of Coling, Mumbai, India, pp. 163–174,2012.*
5. *N. Joshi, H. Darbari, I. Mathur. 2013.HMM Based POS Tagger for Hindi. In Proceedings of International Conference Artificial Intelligence, Soft Computing, CS & IT Proceedings, Vol 3, No 6.*
6. *A. Paul, B. S. Purkayastha and S. Sarkar, "Hidden Markov Model based Part of Speech Tagging for Nepali language,* "International Symposium on Advanced Computing and Communication (ISACC), *Silchar, pp. 149-156, 2015.*
7. *Pravesh KumarDwivedi, Pritendra Kumar Malakar," Hybrid Approach Based POS Tagger for Hindi Language", International Journal of Emerging Technology and Advanced Engineering,2015.*
8. *Antony P J, Dr. Soman K, P," Parts Of Speech Tagging for Indian Languages: A Literature Survey", International Journal of Computer Applications (0975 – 8887) Volume 34– No.8, November 2011.*
9. *Shachi Mall, Umesh Chandra Jaiswal," Innovative Algorithms for Parts of Speech Tagging in Hindi-English Machine Translation, Language", 2015 International Conference on Green Computing and Internet of Things (leGCloT).*
10. *Sanjeev Kumar Sharma and Gurpreet Singh Lehal, "Using Hidden Markov Model to improve the accuracy of a Punjabi POS tagger,"* IEEE International Conference on Computer Science and Automation Engineering, *Shanghai, pp. 697-701, 2011.*
11. *D. Gunasekara, W. V. Welgama and A. R. Weerasinghe, "Hybrid Part of Speech tagger for Sinhala Language,"* Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), *Negombo, pp. 41-48, 2016.*